COMP307/AIML420 INTRODUCTION TO ARTIFICIAL INTELLIGENCE

Reasoning under uncertainty:

Bayesian Networks

Outline

- Review of Bayes Theorem and Naïve Bayes
- Introduction to Bayesian Networks

Review

Product rule: $P(A, B) = P(B) * P(A \mid B) = P(A) * P(B \mid A)$

Bayes theorem:

- Provides a way to calculate the probability of a <u>hypothesis</u> (e.g., **label**) given some <u>evidence</u> (e.g., **feature values**).

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}$$

Naïve Bayes:

- Probabilistic classifier
- Training: count and store <u>priors</u> and <u>likelihoods</u>
- Assumes features are **conditionally independent given the class label**

Review

Product rule: $P(A, B) = P(B) * P(A \mid B) = P(A) * P(B \mid A)$

Bayes theorem:

- Provides a way to calculate the probability of a <u>hypothesis</u> (e.g., **class label**) given some <u>evidence</u> (e.g., **feature values**).

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}$$

Naïve Bayes:

- Probabilistic classifier
- Training: count and store <u>priors</u> and <u>likelihoods</u>
- Assumes features are **conditionally independent given the class label**

Review

- Naïve Bayes:

$$p(c\,|x_1, \cdots, x_p) = \frac{p(c)p(\,x_1, \cdots, x_p|c)}{p(x_1, \cdots, x_p)}$$

class conditional feature independence

$$= \frac{p(c)p(x_1|c) \cdots p(x_p|c)}{p(x_1, \cdots, x_p)}$$

- We don't need denominator to find c with highest probability
  - Use *score*

Review: Naïve Bayes

- In classification we know the **priors** and the **likelihoods** for the training data

- Example:
  - Class C = **sunny** (true/false)
  - Features: season (spring/summer/fall/winter), humidity (high/low)
  - Want to compute determine
    if sunny=true or sunny=false given that $humidity = high,\ season = summer$
  - Probabilities needed:
    - P(season=summer l sunny=true) : count instances where season = summer and class label sunny = true and divide by total number of instances where sunny = true
    - P(season=summer l sunny=false)
    - P(humidity =high l sunny=true)
    - P(humidity =high l sunny=false)
    - P(sunny=true) : count the instances sunny=true and divide by the total number of instances
    - P(sunny=false)

- What is the conditional independence assumption? Is it reasonable?
- Why do we calculate a $score$ instead of the posterior probability $P(A|B)$?

Review: Naïve Bayes

Given an instance x = (**season**=Summer,**humidity**=high), we need to calculate:

$$P(\boldsymbol{sunny} = True \mid \boldsymbol{season} = Summer, \boldsymbol{humidity} = high)$$

$$P(\boldsymbol{sunny} = False \mid \boldsymbol{season} = Summer, \boldsymbol{humidity} = high)$$

Review: Naïve Bayes

Given an instance x = (**season**=Summer,**humidity**=high), we need to calculate:

$$P(\boldsymbol{sunny} = True \mid \boldsymbol{season} = Summer, \boldsymbol{humidity} = high)$$

$$P(\boldsymbol{sunny} = False \mid \boldsymbol{season} = Summer, \boldsymbol{humidity} = high)$$

*

$$P(\boldsymbol{sunny} = True \mid \boldsymbol{season} = Summer, \boldsymbol{humidity} = high) = P(\boldsymbol{season} = Summer, \boldsymbol{humidity} = high \mid \boldsymbol{sunny} = True) * P(\boldsymbol{sunny} = True) / P(\boldsymbol{season} = Summer, \boldsymbol{humidity} = high)$$

Review: Naïve Bayes

Given an instance x = (**season**=Summer,**humidity**=high), we need to calculate:

$$P(\boldsymbol{sunny} = True \mid \boldsymbol{season} = Summer, \boldsymbol{humidity} = high)$$

$$P(\boldsymbol{is\_sunny} = False \mid \boldsymbol{season} = Summer, \boldsymbol{humidity} = high)$$

$$P(\boldsymbol{sunny} = True \mid \boldsymbol{season} = Summer, \boldsymbol{humidity} = high) =$$
$$P(\boldsymbol{season} = Summer, \boldsymbol{humidity} = high \mid \boldsymbol{sunny} = True) * P(\boldsymbol{sunny} = True) / P(\boldsymbol{season} = Summer, \boldsymbol{humidity} = high)$$

$$P(\boldsymbol{season} = Summer \mid \boldsymbol{sunny} = True)\, P(\boldsymbol{humidity} = high \mid \boldsymbol{sunny} = True)$$

Conditional independence assumption. Is it reasonable?

Review: Naïve Bayes

Given an instance x = (**season**=Summer,**humidity**=high), we need to calculate:

$$P(\boldsymbol{sunny} = True \mid \boldsymbol{season} = Summer, \boldsymbol{humidity} = high)$$

$$P(\boldsymbol{is\_sunny} = False \mid \boldsymbol{season} = Summer, \boldsymbol{humidity} = high)$$

$$P(\boldsymbol{sunny} = True \mid \boldsymbol{season} = Summer, \boldsymbol{humidity} = high) =$$
$$P(\boldsymbol{season} = Summer, \boldsymbol{humidity} = high \mid \boldsymbol{sunny} = True) * P(\boldsymbol{sunny} = True) /$$
$$P(\boldsymbol{season} = Summer, \boldsymbol{humidity} = high)$$

calculating P(X) not needed for class label

$$P(\boldsymbol{season} = Summer \mid \boldsymbol{sunny} = True)\, P(\boldsymbol{humidity} = high \mid \boldsymbol{sunny} = True)$$

conditional independence assumption

Review: Naïve Bayes

Given an instance x = (**season**=Summer,**humidity**=high), we need to calculate:

$$P(\boldsymbol{sunny} = true \mid \boldsymbol{season} = summer, \boldsymbol{humidity} = high)$$

$$P(\boldsymbol{sunny} = false \mid \boldsymbol{season} = summer, \boldsymbol{humidity} = high)$$

score$(\boldsymbol{sunny} = true \mid \boldsymbol{season} = summ, \boldsymbol{hum} = high)$=

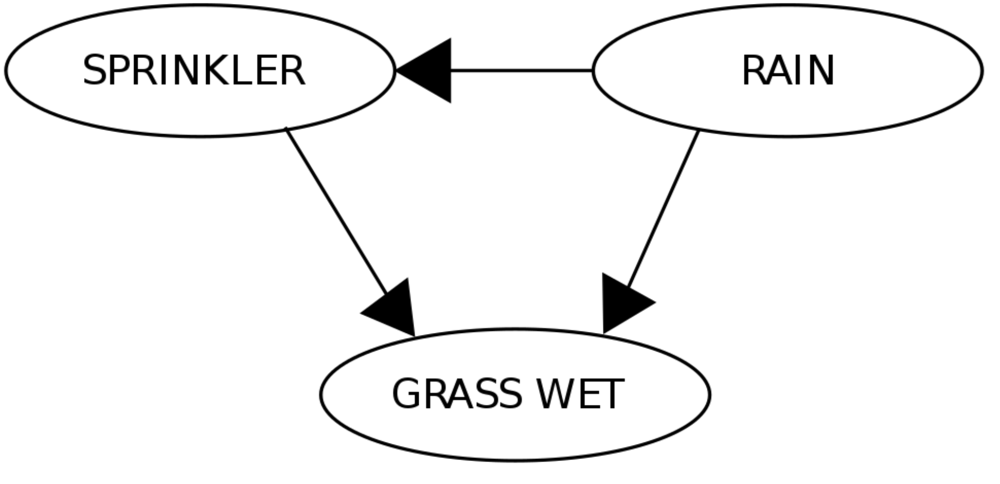
$$P(\boldsymbol{season} = summ \mid \boldsymbol{sunny} = true) *$$
$$P(\boldsymbol{hum} = high \mid \boldsymbol{sunny} = true) *$$
$$P(\boldsymbol{sunny} = true)$$

Bayesian Networks

- **Bayesian networks (BNs)** are a type of **probabilistic graphical model** that represents the joint probability distribution over **a set of random variables** and their **conditional dependencies** using a directed acyclic graph (DAG).

- **Node:** a random variable

- **Edge:** represent *causal* dependencies between nodes



Source: By AnAj - Public Domain

BN Example 1

Given an electric fan, suppose you try to turn it on, but it doesn't spin (not working).

Why is the fan not spinning?

- **Faulty fan**: the fan is broken
- **Faulty plug**: the plug is broken
- A **phone charger** connected to the same plug works well
- *"**Faulty Fan** and **Faulty Plug** are **marginally independent**; however, they become **conditionally dependent, given Fan**."* [1]

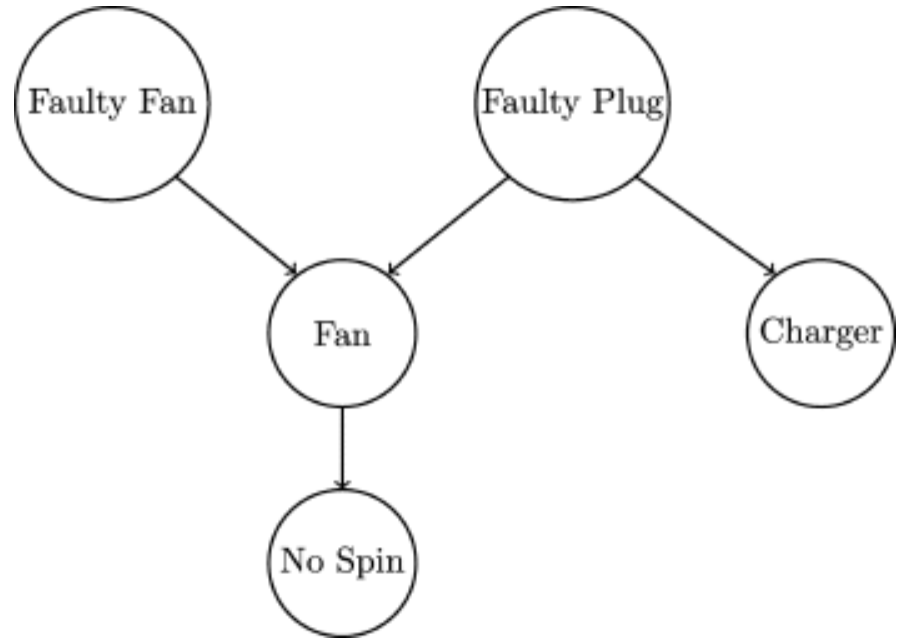


Figure: Simple BN [1]

[1] Yang, Xin-She. *Introduction to algorithms for data mining and machine learning*. Academic press, 2019.

BN Example 1

Given an electric fan, suppose you try to turn it on, but it doesn't spin (not working).

Why is the fan not spinning?

- **Faulty fan**: the fan is broken
- **Faulty plug**: the plug is broken
- A **phone charger** connected to the same plug works well
- *"**Faulty Fan** and **Faulty Plug** are **marginally independent**; however, they become **conditionally dependent, given Fan**."* [1]

- Important concepts:
  - **(marginally) independent:** if we know the (marginal) probability distribution of one variable, it does not affect the probability distribution of the other variable
  - **conditionally independent:** the variables are independent when another variable is known

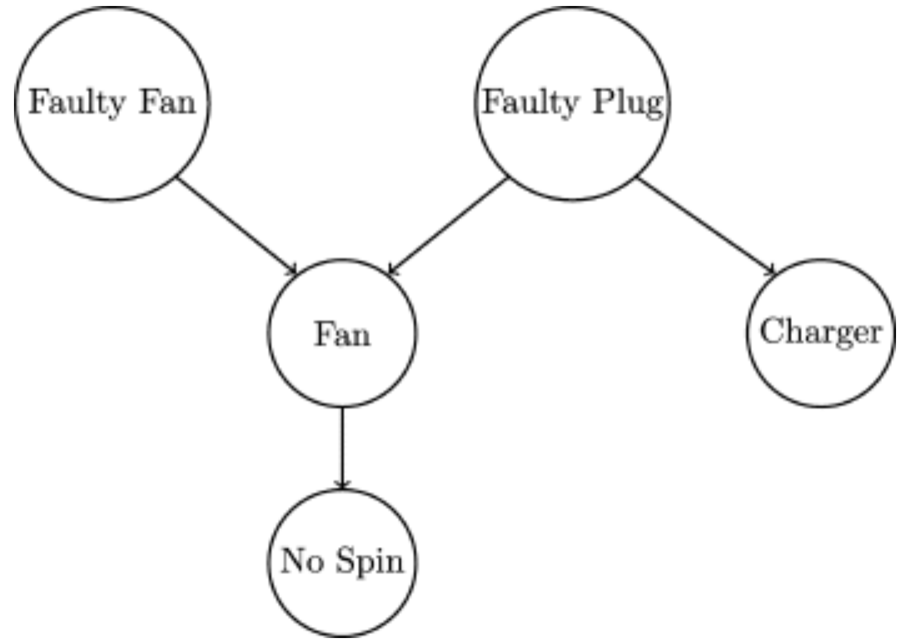


Figure: Simple BN [1]

[1] Yang, Xin-She. *Introduction to algorithms for data mining and machine learning*. Academic press, 2019.

BN Example 1

Given an electric fan, suppose you try to turn it on, but it doesn't spin (not working).

Why is the fan not spinning?

- **Faulty fan**: the fan is broken
- **Faulty plug**: the plug is broken
- A **phone charger** connected to the same plug works well
- *"**Faulty Fan** and **Faulty Plug** are **<u>marginally independent</u>**; however, they become **<u>conditionally dependent, given Fan</u>**."* [1]

- Important concepts:
  - **<u>(marginally) independent:</u>** if we know the (marginal) probability distribution of one variable, it does not affect the probability distribution of the other variable
  - **<u>conditionally independent</u>:** the variables are independent when another variable is known

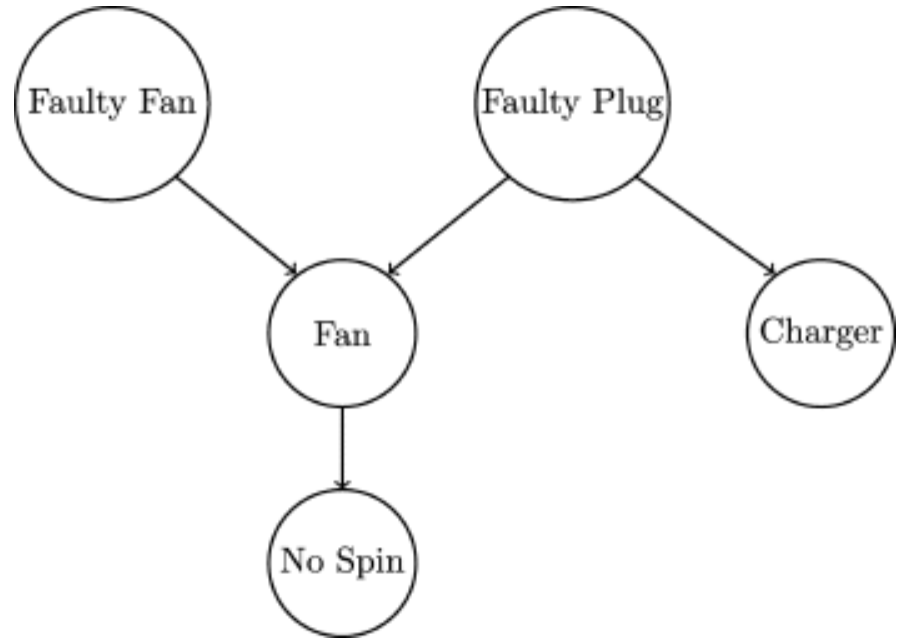


Figure: Simple BN [1]

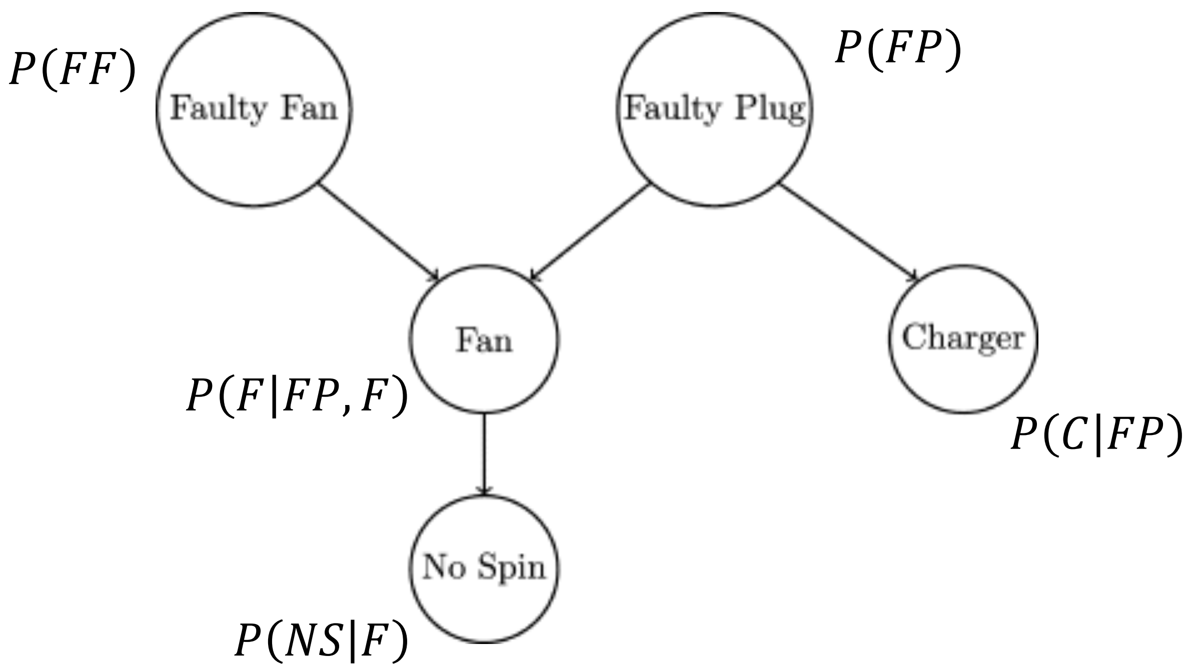
BN models the joint probability distribution of a set of random variables by decomposing it into a product of conditional probabilities

[1] Yang, Xin-She. *Introduction to algorithms for data mining and machine learning*. Academic press, 2019.

BN Example 1

- Work backward from the bottom to obtain:

$$
\begin{aligned}
P(FF, FP, F, C, NS) &= P(NS|\, FF, FP, F, C\,)P(FF, FP, F, C) \\
&= P(NS|F)P(F|FP, F, C)P(FF, FP, C) \\
&= P(NS|F)P(F|FP, F)P(C|FF, FP)P(FF, FP) \\
&= P(NS|F)P(F|FP, F)P(C|FP)P(FF)P(FP)
\end{aligned}
$$



BN Example

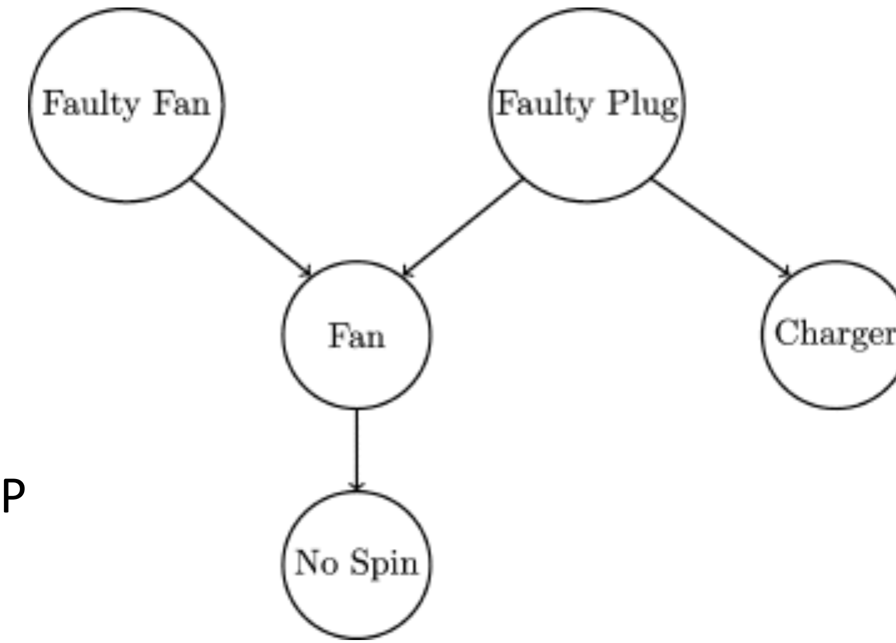
$$
\begin{aligned}
P(FF, FP, F, C, NS) &= P(NS|\, FF, FP, F, C\,)P(FF, FP, F, C) \\
&= P(NS|F)P(F|FP, F, C)P(FF, FP, C) \\
&= P(NS|F)P(F|FP, F)P(C|FF, FP)P(FF, FP) \\
&= P(NS|F)P(F|FP, F)P(C|FP)P(FF)P(FP)
\end{aligned}
$$

- All features have two states. Five features. Joint probability $P(FF, FP, F, C, NS)$ has table with 32 entries (31 sufficient).
- Instead, we have five tables with 2+4+2+1+1 = 10 entries.
  - For example, $P(NS|F)$ has 2 entries; exploit that $P(NS = 0|F) = 1 - P(NS = 1|F)$ so store only one of these.
- Fewer parameters, hence fewer data needed for their estimation. Parameters replaced with structural knowledge.
  - More robust results.

BN Example

Useful rule:

- Given the parents of a node A, the node A is independent of its non-descendants



Which are true?

FF and FP are independent
FF and FP independent given F
F and C are independent
F and C are independent given FP
NS and C are independent
NS and C are independent given FP
NS and C are independent given F

Yes, no, no, yes, no, yes, yes

BN Example 2

Modelling the relationship between **Sprinkler**, **Rain** and **Grass Wet**

- Two events can cause the grass to become wet (**Rain**=True or **Sprinkler**=True)

BN Example 2

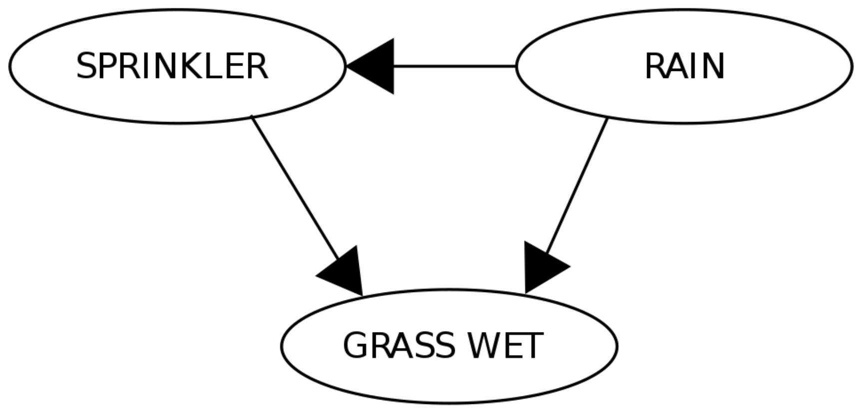
Modelling the relationship between **Sprinkler**, **Rain** and **Grass Wet**

- Two events can cause the grass to become wet (**Rain**=True or **Sprinkler**=True)
- If **Rain**=True, then Sprinkler is unlikely to be True

BN Example 2

Modelling the relationship between **Sprinkler**, **Rain** and **Grass Wet**

- Two events can cause the grass to become wet (**Rain**=True or **Sprinkler**=True)
- If **Rain**=True, then Sprinkler is unlikely to be True



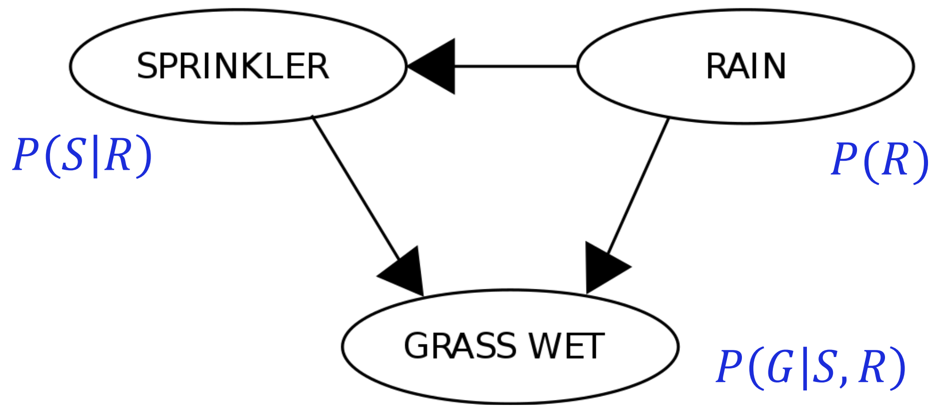
Source: By AnAj - Public Domain

BN Example 2

Modelling the relationship between **Sprinkler**, **Rain** and **Grass Wet**

Let *G = "Grass wet", S = "Sprinkler turned on", and R = "Raining"* .
Joint probability is:

$$P(R,S,G) = P(G|S,R)\,P(S,R) = P(G|S,R)P(S|R)P(R)$$



Source: By AnAj - Public Domain

BN Example 2

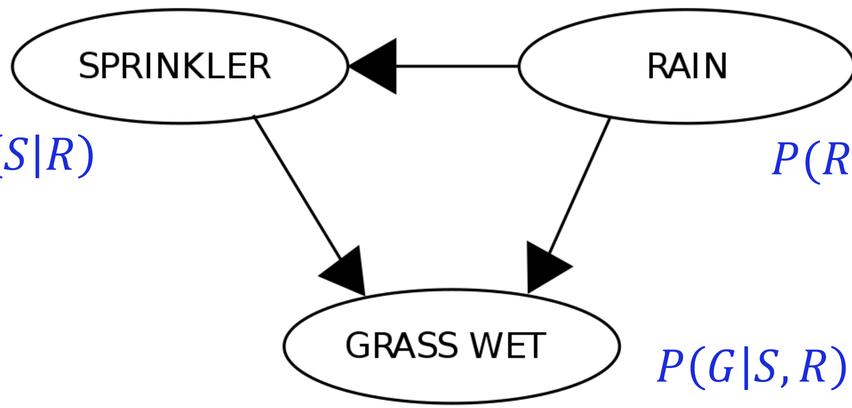
Modelling the relationship between **Sprinkler**, **Rain** and **Grass Wet**

Let *G = "Grass wet", S = "Sprinkler turned on", and R = "Raining"* .
Joint probability is:

$$P(R,S,G) = P(G|S,R)\ P(S,R) = P(G|S,R)P(G|S,R)(S|R)P(R)$$

RAIN	SPRINKLER T	SPRINKLER F
F	0.4	0.6
T	0.01	0.99

$P(S|R)$



$P(R)$

RAIN T	RAIN F
0.2	0.8

$P(G|S,R)$

SPRINKLER	RAIN	GRASS WET T	GRASS WET F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

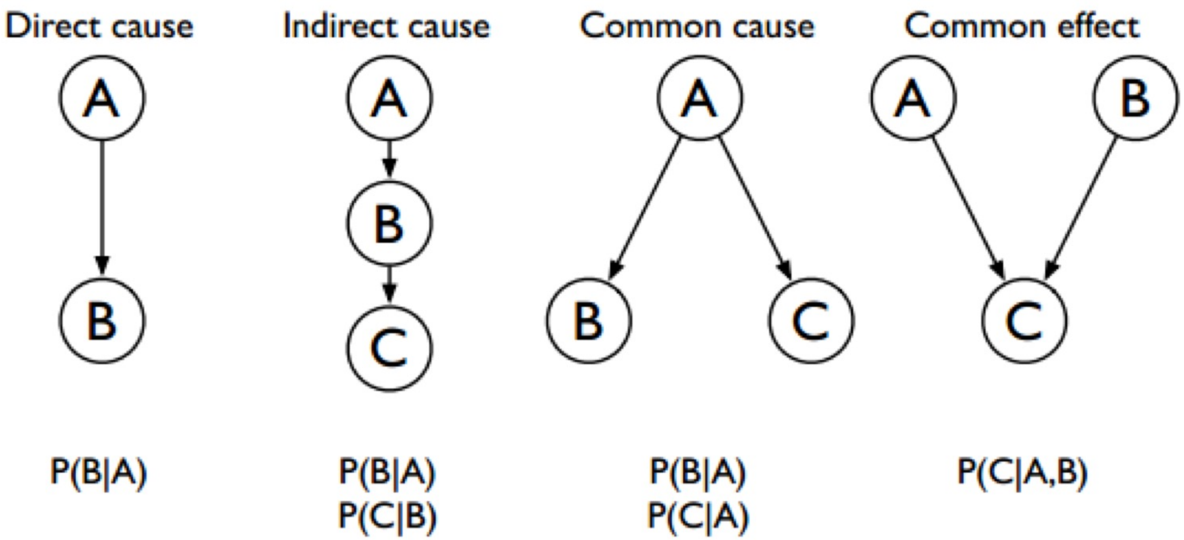
Source: By AnAj - Public Domain

Why Bayesian Networks?

- The model encodes dependencies among all variables
- A BN can be used to learn causal relationships, and hence gain understanding about a problem domain and to predict the consequences of intervention
- The model has both a causal and probabilistic semantics, it is an ideal representation for combining prior knowledge (which often comes in causal form) and data
- Bayesian statistical methods in conjunction with Bayesian networks offer an efficient and principled approach for reduce the overfitting of data. (As it enforces a presumably correct structure.)
- A BN requires typically far fewer parameters and hence less storage than a full joint probability distribution

(In)Dependencies in BN

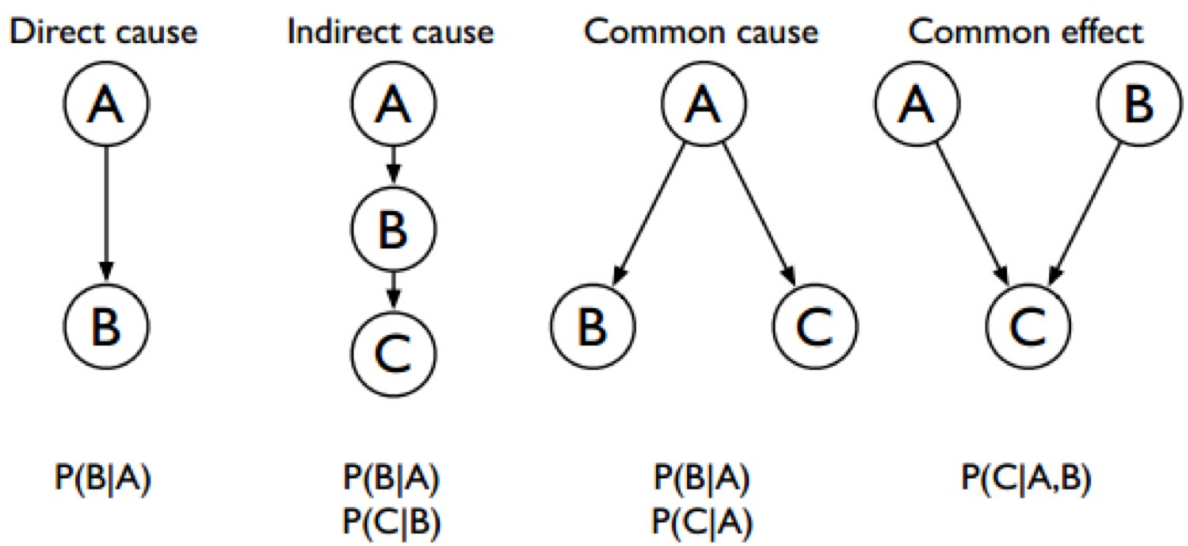
- [**Direct cause**]: A is a direct cause of B
  - A and B are dependent
- [**Indirect cause**]: A direct cause of B, B direct cause of C
  - A and C are independent given B



$$P(C|A) = \sum_n P(C, B_n|A)$$
$$= \sum_n P(C|B_n, A)\, P(B_n|A)$$
$$= \sum_n P(C|B_n)\, P(B_n|A)$$

(In)Dependencies in BN

- [**Common Cause**]: A is a **common direct cause** of B and C
  - B and C are dependent **(if A is not given)**
  - B and C are independent **given A**
- [**Common Effect**]: C is a **common direct effect** of A and B
  - A and B are independent **(if C is not given)**
  - A and B are dependent **given C** ("explaining away")

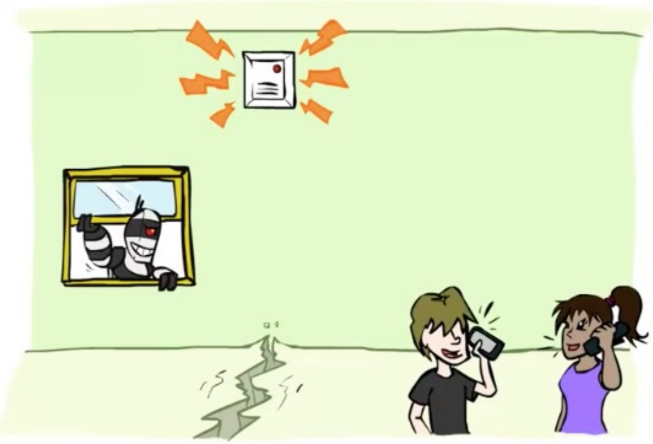


BN Example 3

- Your house has an alarm against burglary
  - The alarm will usually be set off by burglars
  - Sometimes it may also be set off by earthquakes
  - There are two neighbours, John and Mary
  - John and Mary might call you when they hear the alarm
  - They might also call you for other issues without alarm

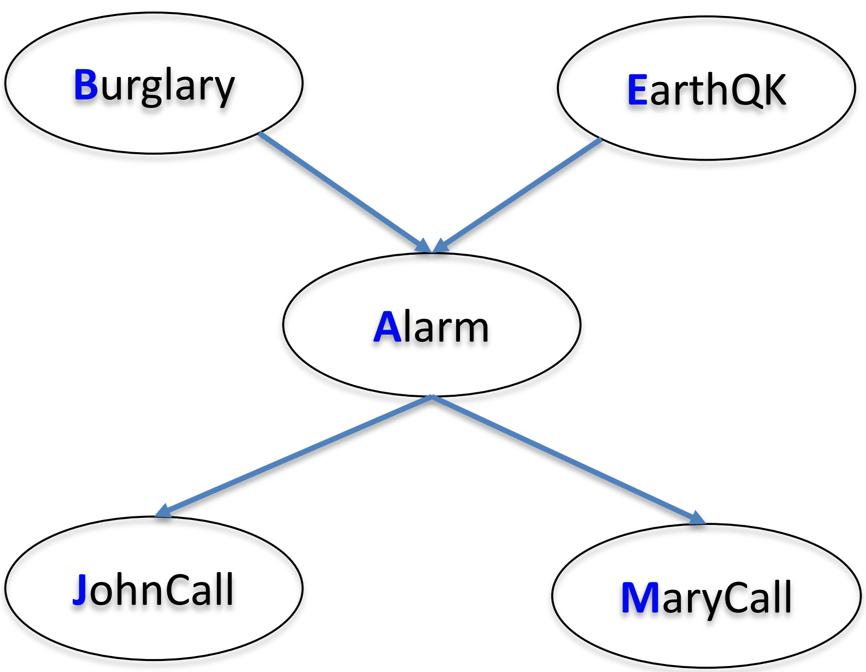
- Variables**:**
  - Burglary, Earthquake, Alarm, JohnCalls, MaryCalls
  - All binary (**true** or **false)**

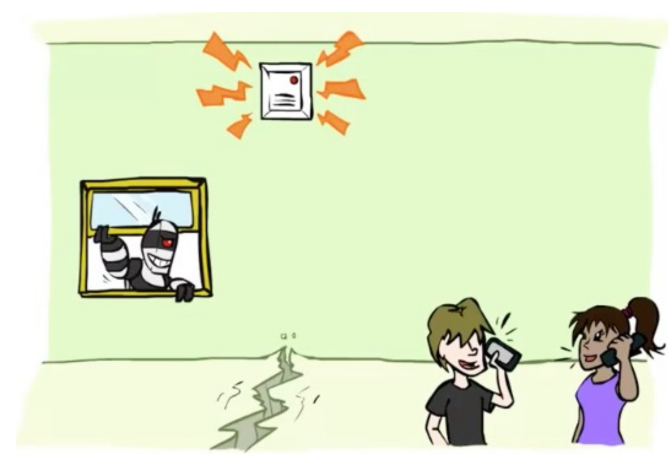
- Relationship between them?
  - Cause -> Effect



Alarm Network

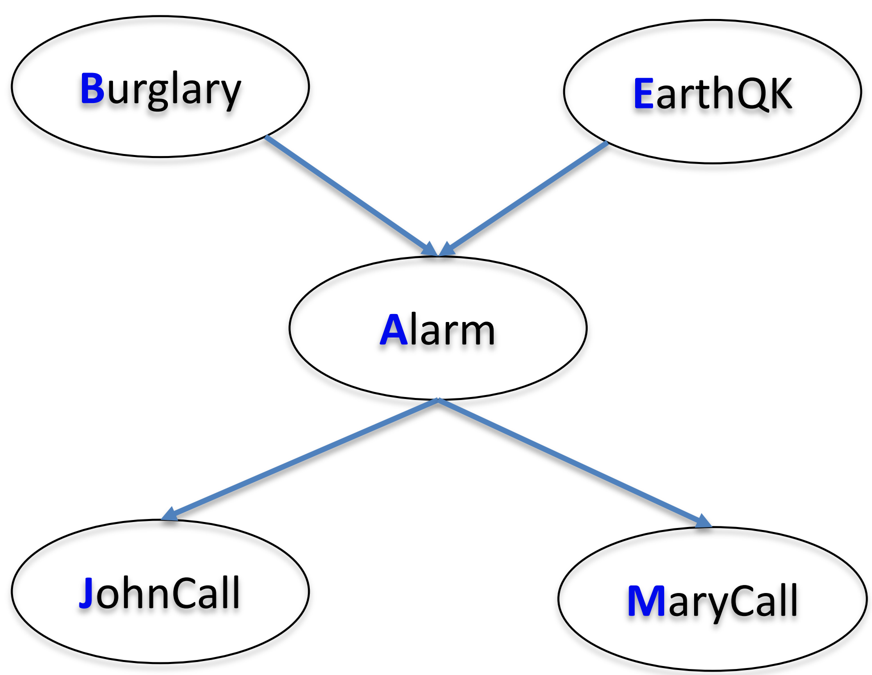
- Domain **causal knowledge (causes and effects)**
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call





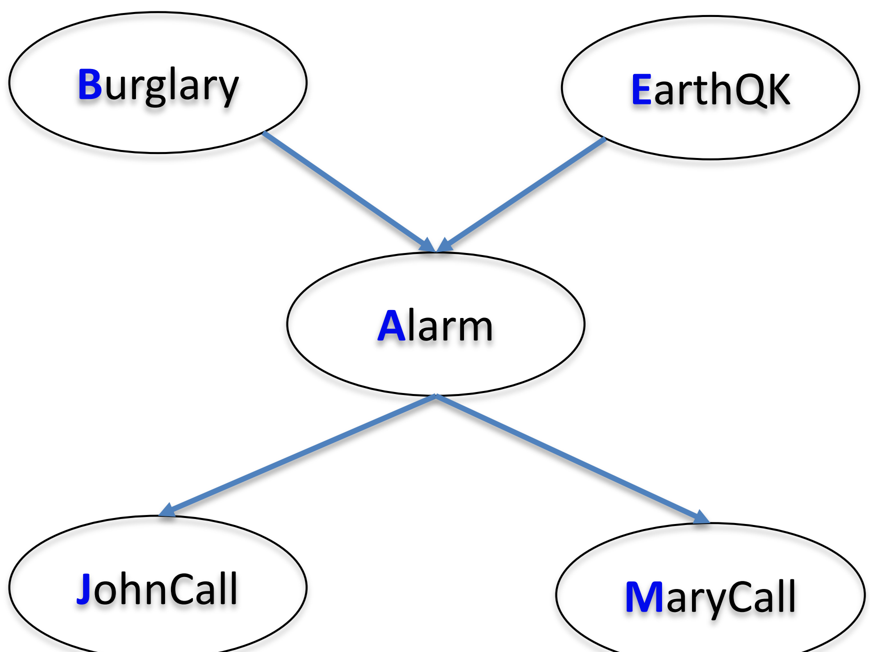
(In)Dependencies in BN

- Recall: given the parents of a node A, the node A is independent of its non-descendants



(In)Dependencies in BN

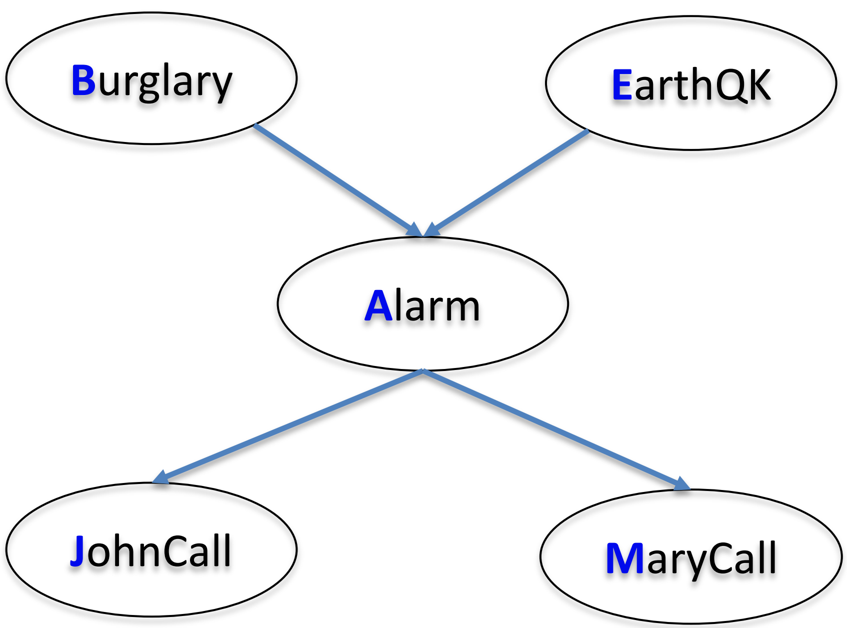
- Which are true?
  - B and E are independent
  - B and E are independent given A
  - B and M are independent
  - B and M are independent given A
  - J and M are independent
  - J and M are independent given A



Factorisation

$$P(M,J,A,E,B\ ) = P(M,J,A|\ E,B)P(E,B)$$
$$= P(M,J,A\ |\ E,B)\ P(E)P(B)$$
$$= P(M,J|A,E,B)P(A|E,B)P(E)P(B)$$
$$= P(M,J|A)P(A|E,B)P(E)P(B)$$
$$= P(M|A)P(J|A)P(A|E,B)P(E)P(B)$$

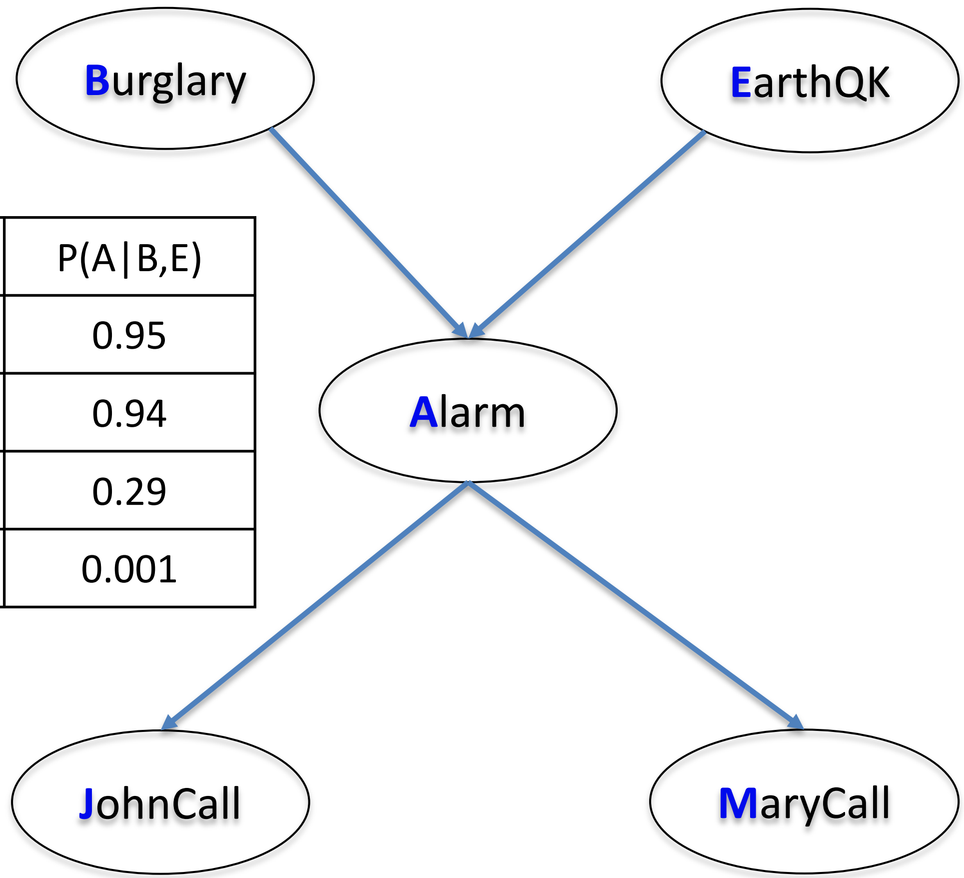
- Each conditional probability is a node.
- Table with 32 entries becomes tables with 2+2+4+1+1=10 entries



Alarm Network

B	P(B)
T	0.001

E	P(E)
T	0.002



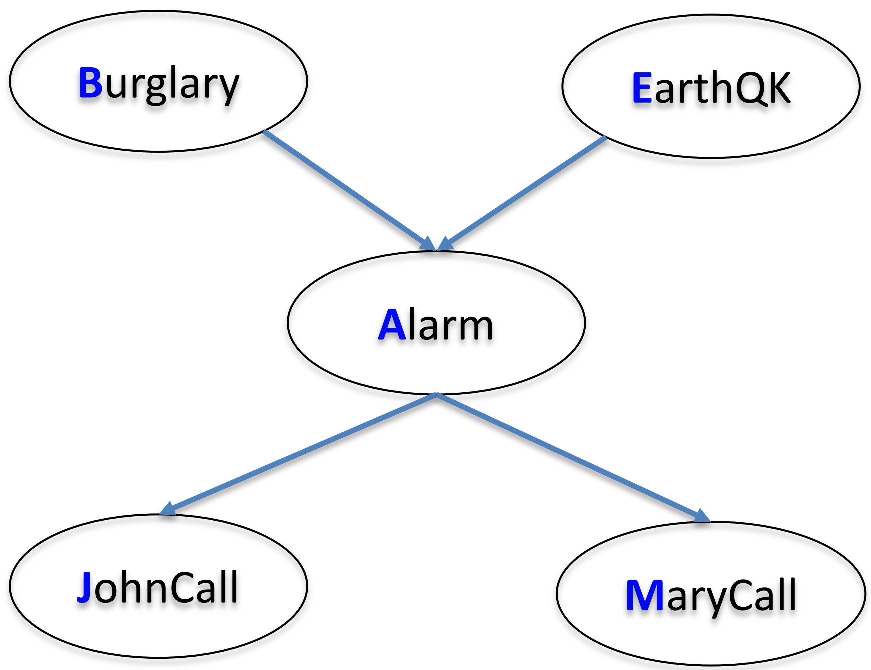
B	E	A	P(A|B,E)
T	T	T	0.95
T	F	T	0.94
F	T	T	0.29
F	F	T	0.001

A	J	P(J|A)
T	T	0.9
F	T	0.05

A	M	P(M|A)
T	T	0.7
F	T	0.01

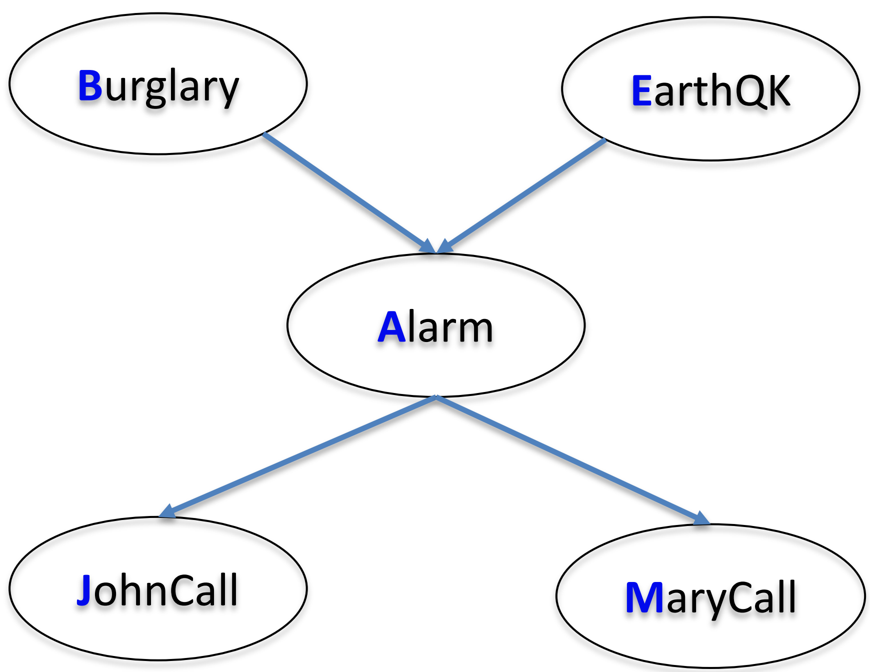
Factorisation

- To factorise according to a Bayesian network: sort the variables so that the causes are always before the effects, e.g., [B, E, A, J, M], then use the rules:
  - $P(X_i \mid X_1, \cdots, X_{i-1}) = P\big(X_i \,\big|\, \text{parents}(X_i), \cdots, X_j, \cdots\big)$
    $= P\big(X_i \,\big|\, \text{parents}(X_i)\big)$
  - $P\big(X_i \,\big|\, X_j\big) = P(X_i)$ if X_i and X_j independent



Factorisation

- We saw repeated application of the product rule gives
  $P(M, J, A, E, B\ ) = P(M|A)P(J|A)P(A|E, B)P(E)P(B)$
- The joint probability distribution over all variables in the network can be represented as a product of the conditional probabilities of each variable given its parents:
  $P(X_1, \cdots, X_n) = P(X_n|\text{parents}(X_n))P(X_{n-1}|\text{parents}(X_{n-1}))\cdots$



Summary

- Conditionally independent given class label (NB)
- Bayes Net = Topology (graph) + Local Conditional Probabilities
- Factorisation

Coming up next…

- More on Bayesian Networks (Build a BN, # free parameters, …)